WERNER BRECHT – BRIGITTE DOMKE – HANS-DIETER LUTZ – DIETER METZING

Konstruktionsprinzipien für ein Lexikon in der maschinellen Sprachverarbeitung

In der Abteilung "Linguistische Datenverarbeitung" (LDV) des Instituts für deutsche Sprache wird ein "Informationssystem auf linguistischer Basis" (ISLIB) entwickelt. An dieses System werden eine ganze Reihe von Anforderungen gestellt:

1. Es ist ein Dialog-System und der Dialog erfolgt in natürlicher Sprache, und zwar in Deutsch.
2. Das System hat problemlösende Eigenschaften.
3. Es ist ein Experimentalsystem. Das bedeutet, daß bei der Bearbeitung von Teilen von ISLIB durchaus verschiedene Ansätze verfolgt werden können.

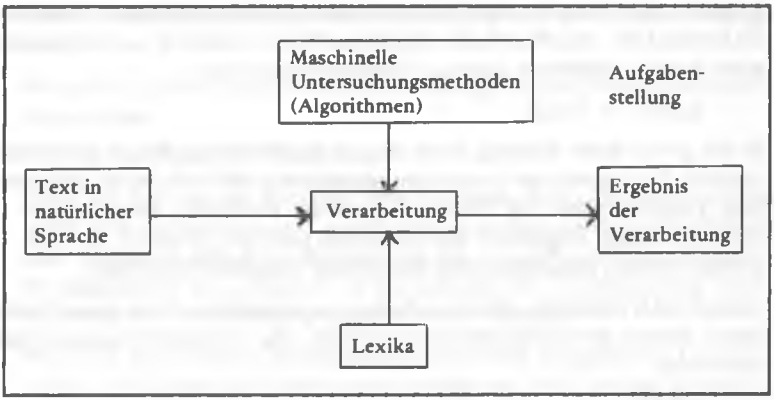
Es ist in diesem Rahmen nicht möglich, den gegenwärtigen Stand der Arbeiten an ISLIB in allen Einzelheiten darzustellen. Hingegen scheint es uns sinnvoll, den Begriff des Lexikons in den Vordergrund zu stellen; dieser Begriff spielt einmal in ISLIB, zum andern auch in nahezu allen Teilen der maschinellen Textverarbeitung, wie

statistische Linguistik,
Indexierung,
maschinelle Sprachanalyse,
maschinelle Übersetzung,
Informationssysteme,
Computer-unterstützter Unterricht,
usw.

eine große Rolle.

Es sei an dieser Stelle bemerkt, daß aus den folgenden Darlegungen nicht gefolgert werden darf, daß diese Betrachtungen bereits in vollem Umfange in ISLIB eingebracht sind.

Die Rolle des Lexikons in der maschinellen Sprachbearbeitung können wir uns in sehr vereinfachter Form an dem folgenden Schaubild klarmachen.



Dieses Bild ist folgendermaßen zu interpretieren:

Gegeben ist eine bestimmte Aufgabenstellung, z.B. eine maschinelle Übersetzung von einer natürlichen Sprache in eine andere. In dieser Aufgabenstellung ist ein natürlich-sprachlicher Text gegeben, der verarbeitet werden soll. Dazu werden benötigt:

(a) Untersuchungsmethoden, die von einer Maschine ausgeführt werden können, also algorithmisiert vorliegen müssen, und

(b) Lexika, denen die Maschine ihr aufgaben-spezifisches Fachwissen entnimmt.

Aus dem Text, den Untersuchungsmethoden und den Lexika gewinnt die Maschine das Verarbeitungsergebnis.

Es ist nun offensichtlich, daß die gegebene Aufgabenstellung die verwendeten Lexika beeinflußt. Denn wenn jemand versuchen will, maschinell Shakespeare zu übersetzen, wird er bestimmt ein anderes Lexikon benötigen als jemand, der lediglich auszählen will, wie oft in den Werken von Karl May das Wort *Hinterlader* vorkommt.

Wir wollen nun im folgenden einmal systematisch untersuchen, wie sich mit immer anspruchsvoller, mit immer komplexer werdenden Aufgabenstellungen die Struktur der verwendeten Lexika ändert.

Beginnen wollen wir mit der wohl elementarsten Untersuchung eines Textes überhaupt, nämlich dem Zählen der Wortformen durch Zählen der Wortlücken. Wider Erwarten benötigt man hier bereits ein Lexikon, denn man muß ja irgendwo "nachschauen", ob ein bestimmtes Zeichen eine Wortlücke ist oder nicht; und nachschauen tut man bekanntlich in einem

Lexikon. Allerdings ist dieses Lexikon sehr einfach aufgebaut. Wenn wir als Symbol für eine Wortlücke ein oben offenes Rechteck ␣ verwenden, sieht das hier benötigte Lexikon folgendermaßen aus:

LEX1 : = {␣}.

Es hat genau einen Eintrag, kann also als einelementige Menge aufgefaßt werden. Die zugehörige Untersuchungsmethode läßt sich leicht beschreiben: Jedes Graphem des Textes wird mit der Wortlücke, also mit dem Lexikoneintrag, verglichen. Bei Gleichheit wird ein Zählwerk um eine Einheit erhöht, ansonsten wird das nächste Graphem verglichen.

Immer noch elementar ist die Aufgabe, die Anzahl der Sätze eines Textes durch Zählen der Satzzeichen festzustellen. Das verwendete Lexikon hat die Form:

LEX2 : = {. , ; ! ? ...}.

Jedes Graphem des zu untersuchenden Textes wird mit jedem Zeichen aus LEX2 verglichen. Bei Gleichheit wird ein Zähler erhöht, sonst wird das nächste Graphem verglichen.

Man kann sich jetzt eine ganze Reihe von Aufgabenstellungen ausdenken, bei denen die verwendeten Lexika genau so aufgebaut sind wie LEX1 und LEX2, nur mit anderen Zeichen (z.B. ganzen Wörtern) und anderem Umfang.

So könnte ein Lexikon

LEX3 : = {A,B,C,...Z,a,b,c,...,z}

zur Alphabetisierung eines Textes verwendet werden.

Was alle diese Lexika gemeinsam haben, erkennt man sofort, wenn man sich anschaut, wie ein Eintrag in einem traditionellen Wörterbuch aussieht. Betrachten wir das folgende Beispiel aus dem DUDEN (Rechtschreibung, 17. Auflage, Mannheim 1973):

> Trip engl. (Ausflug, Reise; Rauschzustand durch Drogeneinwirkung, auch: die dafür benötigte Dosis) m; -s, -s

Dieser Lexikoneintrag besteht aus einem Identifikationsteil, nämlich dem Wort *Trip,* und einem Beschreibungsteil, in dem Aussagen über Aussprache, Herkunft, Bedeutung und morphologische Merkmale gemacht werden.

Allgemein können wir sagen, daß ein Lexikoneintrag immer zweiteilig ist. Er hat die Form

(a, b).

Dabei ist a der Identifikationsteil des Eintrags und b der Beschreibungsteil.

Das ganze Lexikon kann dann als eine Menge solcher Einträge verstanden werden:

$$L := \{(a,b)_i \mid i = 1,2,3,\ldots,n\}.$$

Wenn wir hier von der Struktur eines Lexikons reden, meinen wir damit nicht, daß die Einträge in L nach ihrem Identifikationsteil alphabetisch oder rückläufig alphabetisch oder nach Häufigkeiten oder sonst irgendwie angeordnet sind, sondern wir beschreiben die Strukturierung eines Lexikons über den Beschreibungsteil seiner Einträge.

Wenn wir nach diesen Gesichtspunkten die oben eingeführten Lexika LEX1, LEX2 und LEX3 betrachten, stellen wir fest, daß der Beschreibungsteil der Einträge fehlt. Genauer müssen wir hier sagen, daß der Beschreibungsteil leer ist und deshalb weggelassen werden kann. Wir wollen alle Lexika dieser Art in einer Klasse zusammenfassen, nennen sie die Lexikonklasse 1 und können somit formulieren: Bei einem Lexikon der Klasse 1 ist der Beschreibungsteil der Einträge leer.

Der nächst komplexe Fall besteht nun darin, daß die Lexikoneinträge zwar einen Beschreibungsteil besitzen, bei diesem aber keinerlei Struktur vorhanden ist. Ein solches Lexikon liegt z.B. vor, wenn ein Index erstellt worden ist, der alle Wörter eines gegebenen Textes und die zugehörigen Stellungsangaben für diese Wörter enthält.

$$\text{LEX4} := \{(\text{Wort, angeordnete Zahlen})_i,\ i = 1,2,3,\ldots,n\}.$$

Lexika dieser Art bezeichnen wir systematisch als Lexika der Klasse 2.

Wesentlich interessanter sind ja nun die Lexika, bei denen der Beschreibungsteil der Einträge tatsächlich eine Struktur aufweist. Und hier ist eine völlig unterschiedliche Komplexität möglich.

Betrachten wir dazu als Beispiel einen Eintrag in einem Lexikon für eine morpho-syntaktisch-semantische Analyse des Deutschen. Es handelt sich um die Verbform *stiegen* und die zugehörige Beschreibung.

```
(STIEGEN
MORPH
(VERB NF STEIGEN PN(4 6) MOD IND TEMP VE DIA AKT VPR (AN –))
SEM1
(V AZ VTR (NOM (N STV)
                (DEVERB (H V)
                        ABSTR
                        (QUANTIFIZIERBAR)
                        GGS
                        (BRA (GEWAESSER)
                             WERTPAPIER)))
   VKAUS
   (PRAEP+N (PRAEP ((DURCH AKK)
                    (MIT DAT)
                    (ANLAESSLICH GEN))
                   N
                   (DEVERB (H V)
                           KKR
                           (NATURKRAFT)
                           ABSTR
                           (AGA))))
   VBERHK
   (PRAEP+N (PRAEP ((VON DAT))
                   N
                   (ABSTR (MASSEINHEIT (ZAHL WAEHRUNG)))))
   VBERZL
   (PRAEP+N (PRAEP ((AUF AKK))
                   N
                   (ABSTR (MASSEINHEIT (ZAHL WAEHRUNG)))))
   VBERS
   (PRAEP+N (PRAEP ((UM AKK))
                   N
                   (ABSTR (MASSEINHEIT (ZAHL WAEHRUNG))))))
SEM2
(H AZ HTR (NOM (N STV)
               (P TIER))
   HBERZL
   (PRAEP+N (PRAEP ((IN AKK)
                    (ZU DAT)
                    (AUF AKK))
                   N
                   (GGS (BRA FAHRZEUG)
                        P TIER)))))
```

Zu dem Identifikationsteil STIEGEN gehört ein Beschreibungsteil, der aus einer morphosyntaktischen, einer ersten semantischen und einer zweiten semantischen Beschreibung besteht.

Die morphosyntaktische Beschreibung bedeutet:
STIEGEN gehört zur Wortklasse VERB. Die Normalform NF ist der Infinitiv STEIGEN. STIEGEN kann erste und dritte Person Plural Indikativ Imperfekt Aktiv sein. Als Verbpräfix kann z.B. AN vorkommen.

Betrachten wir jetzt den semantischen Teil der Beschreibung von STIEGEN. Da unterscheidet sich die semantische Relation VORGANG (V) von der semantischen Relation HANDLUNG (H) durch die unterschiedliche Zuordnung des regierenden Nomen an der mit VTR (Vorgangsträger) bzw. HTR (Handlungsträger) bezeichneten Argumentstelle zu unterschiedlichen semantischen Nomenklassen.
STIEGEN gehört zu VORGANG (V), wenn das regierende Nomen an der Stelle von VTR der Klasse der DEVERBATIVA oder QUANTIFIZIERBAREN ABSTRAKTA oder der GEGENSTÄNDE MIT BETONUNG RÄUMLICHER AUSDEHNUNG zugeordnet werden kann.
STIEGEN gehört zu HANDLUNG (H), wenn das regierende Nomen an der Stelle von HTR der Klasse der PERSONEN oder TIERE zugeordnet werden kann.

Betrachten wir dazu zwei Beispiele:

1. *Die Einkommen der Sekretärinnen und Kontoristinnen stiegen im letzten Jahr um fünf Prozent.*
2. *Die Minister stiegen in die Fahrzeuge.*

Der erste Satz wird der semantischen Relation VORGANG zugeordnet, da *Einkommen* ein QUANTIFIZIERBARES ABSTRAKTUM ist. Der zweite Satz wird HANDLUNG zugeordnet, da *Minister* zur Klasse der PERSONEN gehört.

Lexika der hier beschriebenen Art bezeichnen wir als Lexika der Klasse 3. Formal ist ihr Identifikationsteil unstrukturiert, jedoch betrachtet die Beschreibung den Identifikationsteil nicht mehr isoliert, sondern bereits in bestimmten Umgebungen.

Wir wollen noch anmerken, daß das eben besprochene Beispiel nicht aus der Luft gegriffen ist. Es ist in ISLIB bereits realisiert und erlaubt dort einerseits die morphosyntaktische Analyse eines Eingabesatzes mithilfe einer 'Grammatik', die in einer Netzwerksprache formuliert ist, und andererseits die Monosemierung von Eingabesätzen, bei denen von vornherein nicht klar ist, welchem Sachbereich sie entstammen.

Dieses Beispiel, so komplex es auf den ersten Blick erscheint, wird in der Praxis der LDV noch erweitert um

(a) die Beschreibung von Paraphrasenrelationen. Darunter verstehen

wir synonyme Relationen, antonyme Relationen, inverse Relationen und mehr;

(b) Transformationsregeln, die Deverbativa, Deadjektiva usw. in ihre Bezugsformen überführen.

Es wird jetzt ganz deutlich, daß die Aufgabenstellung die Komplexität der Lexikonstruktur festlegt. Denn wenn nur eine morphologische Analyse deutscher Wortformen durchgeführt werden soll, fallen 95% der in dem Beispiel genannten Merkmale weg. Oder wenn man Texte analysieren will, in denen nur von Handlungen die Rede ist, kann der gesamte Beschreibungsteil, der auf Vorgänge bezogen ist, wegfallen.

Gehen wir jetzt einmal davon aus, daß für die Aufgabenstellung einer automatischen Informationserschließung ein Lexikon der Klasse 3 vorliegt. Dann stellt sich auf dieser Ebene ein Text als eine Folge von Aussagen dar, die mithilfe des Lexikons und der zugehörigen Algorithmen gewonnen werden. Beziehungen zwischen diesen Aussagen können temporär hergestellt werden, indem man bestimmte Verfahren auf eine zuvor nicht geordnete Menge von Aussagen anwendet. Oder aber die Relationierung der Aussagen erfolgt nicht temporär dadurch, daß jede eingegebene Aussage mit Teilen von schon bekannten Aussagen eventuell mit ganzen bereits bekannten Aussagen in einem Netz verknüpft wird.

Die Anwendbarkeit solcher Verfahren in dem Informationssystem ISLIB wird zur Zeit geprüft. Ist eine Informationserschließung wiederum eingebettet in die Aufgabenstellung, z.B. mit einem Computer einen natürlichsprachlichen Dialog zu führen, dann ergibt sich folgendes Bild:

1. Es werden komplexere, und zwar interaktive Algorithmen gebraucht.
2. Das bisher verwendete Lexikon der Klasse 3 reicht nicht aus. Denn daß ein Text aus einer Folge von Aussagen besteht, genügt nicht, um inhaltliche Relationen zwischen diesen Aussagen formulieren zu können.

Um eine solche Relationierung der Aussagen durchzuführen, benötigen wir eine neue Komponente, die als ein Lexikon aufgefaßt werden kann, dessen Identifikationsteil eines Eintrages Aussagen in kanonischen Formulierungen darstellt, während der Beschreibungsteil Beziehungen zwischen Aussagen bzw. zwischen Teilen der Aussagen herstellt.

In der Praxis des Informationssystems ISLIB erfolgt diese Herstellung der Beziehungen zwischen Aussagen durch Algorithmen.

Zusammenfassend läßt sich mit Worten aus der Lexikon-Terminologie festhalten:

Eine Aussagenrelationierung dieser Art besteht aus einem Identifikationsteil, der durch eine Textaussage realisiert ist, und einem Beschreibungsteil, der aus Operationen besteht, die auf den Identifikationsteil zugreifen.

Damit erhalten wir Lexika der Klasse 4. Diese unterscheiden sich von denen der Klasse 3 durch

1. einen dynamischen Beschreibungsteil der Einträge und
2. aufgabenspezifische Interpretationen innerhalb des Beschreibungsteils der Einträge.

Eine weitere fünfte Klasse von Lexika, auf die hier nur ganz kurz eingegangen werden kann, wird dann benötigt, wenn z.B. innerhalb eines natürlich-sprachlichen Dialogs zwischen einem Menschen und einer Maschine die Möglichkeit eröffnet werden soll, nach Beziehungen, z.B. bezüglich Zeit, Ort oder Modalität zu fragen, die zwischen den in der Klasse 4 vorhandenen Aussagenrelationierungen bestehen.

Ein derartiges Lexikon hat als Identifikationsteil der Einträge diejenigen Algorithmen, die als Beschreibungsteile in den Lexika der Klasse 4 vorkommen. Die Beschreibungsteile seiner Einträge enthalten Methoden, die diese Algorithmen in ihrem Zusammenwirken steuern und somit eine Relationierung dieser Algorithmen leisten.

Wenn man nun ein Lexikon für die maschinelle Sprachverarbeitung mit einem der traditionellen Lexika, wie z.B. dem Duden, Wahrig, Mackensen, Brockhaus, Lexikon der deutschen Gegenwartssprache und ähnlichem vergleicht, so fällt auf, daß jeder Eintrag beider Lexikenarten in der Form

(Identifikationsteil, Beschreibungsteil)

vorliegt, auch wenn man leicht feststellen kann, daß im Identifikationsteil eines Eintrages in traditionellen Lexika meistens Merkmale vorhanden sind, die strenggenommen in den Beschreibungsteil gehören, so z.B. Angabe über Aussprache, über Silbengrenzen, Fugenelemente usw.

Neben dieser Gemeinsamkeit gibt es jedoch eine Reihe von Unterschieden.

1. Traditionelle Lexika sind nicht so streng formal aufgebaut wie Computer-Lexika.
2. Bei den Lexika für eine maschinelle Sprachverarbeitung müssen sich die bei den Beschreibungen verwendeten Kategorien in konsistenter Weise auf die algorithmisierten Verarbeitungsprozeduren beziehen.

3. Bei den traditionellen Lexika besteht der Beschreibungsteil stets aus Kategorien, während bei den Computer-Lexika dort auch konkrete Algorithmen stehen können.

Diese Gegenüberstellung ist nun in keiner Weise als eine Kritik der traditionellen Lexika aufzufassen, weil ja die Aufgabenstellungen, in der die Lexika verwendet werden, völlig verschieden sind. Computer-Lexika verlangen aufgrund der maschinenrelevanten formalen Methoden einen diesen Methoden entsprechenden Aufbau, während bei der Erstellung traditioneller Lexika davon ausgegangen werden kann, daß sie einen menschlichen Benutzer haben werden, der diesem formalen Zwang nicht unterliegt.